W290–W295 *Nucleic Acids Research, 2014, Vol. 42, Web Server issue* *Published online 22 May 2014*
doi: 10.1093/nar/gku437

PredHS: a web server for predicting protein–protein interaction hot spots by using structural neighborhood properties

Lei Deng[1,2], Qiangfeng Cliff Zhang[3], Zhigang Chen[1], Yang Meng[1], Jihong Guan[2,*] and Shuigeng Zhou[4,*]

[1]School of Software, Central South University, Changsha 410075, China, [2]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China, [3]Department of Biochemistry and Molecular Biophysics and Center for Computational Biology and Bioinformatics, Columbia University, New York 10032, USA and [4]Shanghai Key Lab of Intelligent Information Processing and School of Computer Science, Fudan University, Shanghai 200433, China

Received March 10, 2014; Revised April 28, 2014; Accepted May 3, 2014

ABSTRACT

Identifying specific hot spot residues that contribute significantly to the affinity and specificity of protein interactions is a problem of the utmost importance. We present an interactive web server, PredHS, which is based on an effective structure-based hot spot prediction method. The PredHS prediction method integrates many novel structural and energetic features with two types of structural neighborhoods (Euclidian and Voronoi), and combines random forest and sequential backward elimination algorithms to select an optimal subset of features. PredHS achieved the highest performance identifying hot spots compared with other state-of-the-art methods, as benchmarked by using an independent experimentally verified dataset. The input to PredHS is protein structures in the PDB format with at least two chains that form interfaces. Users can visualize their predictions in an interactive 3D viewer and download the results as text files. PredHS is available at http://www.predhs.org.

INTRODUCTION

Studies of molecular mechanisms for protein–protein interactions revealed that usually only a small subset of binding interfaces named hot spots account for the majority of binding free energy and are actually critical for stability and function of protein association (1). Identifying and understanding hot spots and their mechanisms on a large scale would have significant implications for practical applications including drug discovery (2) and protein design. Experimentally determined hot spots from alanine scanning mutagenesis experiments have been deposited in Alanine Scanning Energetics Database (ASEdb, (3)). Binding Interface Database (BID) presents experimentally verified hot spots at interfaces collected from literatures (4). However, the number of experimentally determined hot spots deposited in these databases is very limited since experimental techniques to identify hot spots are often labor intensive and expensive. Computational prediction of hot spots has become a practical alternative.

Current approaches for predicting hot spots can be classified roughly into three categories: (i) molecular dynamics (MD) simulations can simulate alanine substitutions and estimate the induced changes in binding free energy ($\Delta\Delta G$) at the atomic level. Some MD-based methods are successful to predict hot spots from protein interfaces (5–8); (ii) knowledge-based methods rely on empirically calibrated free energy functions, which include terms such as van der Waals and electrostatic interactions, hydrogen bonds and solvation energy, providing an alternative way to predict hot spots with much less computation. FOLDEF (9) and Robetta (10) belong to this group and were developed for the fast estimation of mutational free energy changes of a protein for hot spot identification; (iii) machine-learning methods, such as neural networks (11), decision trees (12), support vector machines (13–15), Bayesian networks (16,17), minimum cut trees (18) and random forests (19), have also been applied to detect hot spots in recent years. What's more, several hot spot databases, including HotRegion (20), HotSprint (21) and PCRPi-DB (17), were built based on computational methods.

Although substantial progress has been made, there is significant room for the improvement of protein hot spot prediction. For example, MD-based methods are not ap-

*To whom correspondence should be addressed. Tel: +86 21 55664967; Fax: +86 21 55664298; Email: sgzhou@fudan.edu.cn
Correspondence may also be addressed to Jihong Guan. Tel: +86 21 65109899; Fax: +86 21 65988292; Email: jhguan@tongji.edu.cn

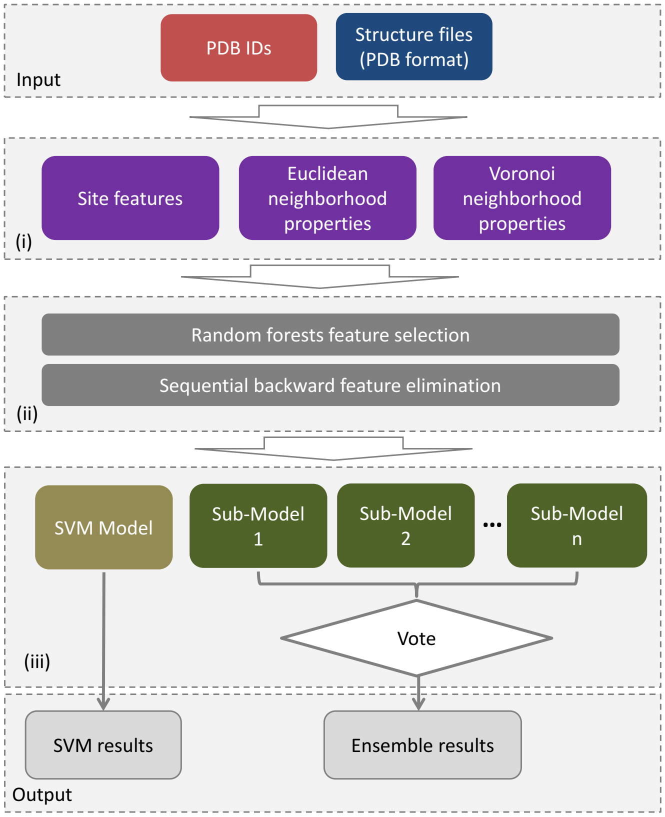
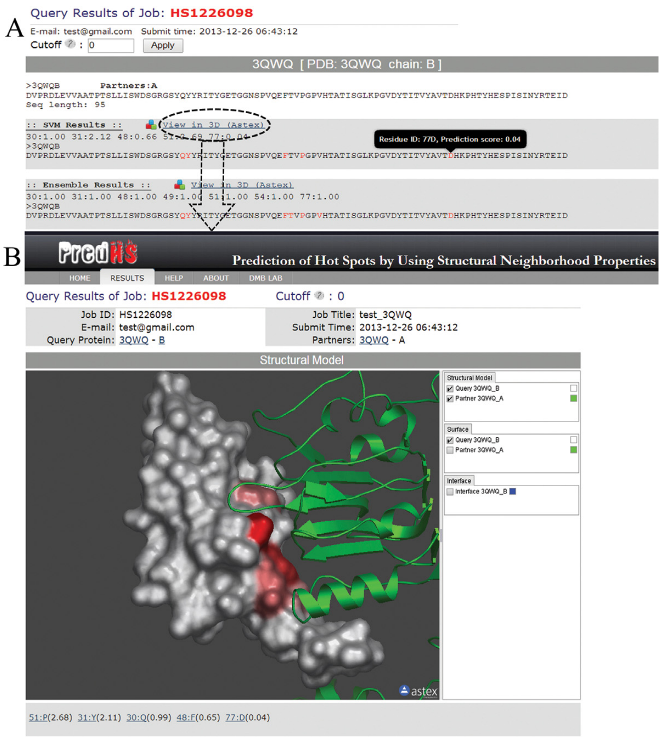

PDB IDs
Structure files (PDB format)
Input
Site features
Euclidean neighborhood properties
Voronoi neighborhood properties
(i)
Random forests feature selection
Sequential backward feature elimination
(ii)
SVM Model
Sub-Model 1
Sub-Model 2
...
Sub-Model n
Vote
(iii)
SVM results
Ensemble results
Output

Figure 1. Flow chart of the PredHS web server. Input to the PredHS server can be protein structure files in PDB format or a list of PDB codes. After job submission, the server invokes three main component processes: (i) feature extraction: a set of 108 site features, 108 Euclidean neighborhood features and 108 Voronoi neighborhood features are extracted; (ii) feature selection: in the training process, a subset of 38 optimal features are selected by using a random forest algorithm and a sequential backward elimination method, these selected features are used for prediction and (iii) prediction models: PredHS-SVM and PredHS-Ensemble, where PredHS-Ensemble is an ensemble of *n* sub-models using a bootstrap resampling method to generate subsets. Finally, two groups of results are obtained (SVM results and Ensemble results), corresponding to the two predictors.

plicable for large-scale studies due to high computational cost. Knowledge-based methods are computationally much faster and reported results appear comparable to those from MD-based simulations (10). But the overall performance of these two groups of methods was inferior to machine-learning methods especially in the measure of recall (22). Machine-learning methods typically depend on the recognition of differences in features including physicochemical properties, evolutionary conservation and solvent accessible area. But specific biological properties for precisely identifying hot spots are often not fully exploited and the performance of the existing methods remains unsatisfactory. Moreover, the number of interacting hot spots of a protein is usually much smaller than the number of energetically unimportant interface residues. Existing methods usually have much higher specificity but rather lower recall since most classification algorithms tend to predict test samples as the majority class and may ignore the minority class when trained on the imbalanced data.



Query Results of Job: HS1226098
E-mail: test@gmail.com Submit time: 2013-12-26 06:43:12
Cutoff : 0 Apply

3QWQ [PDB: 3QWQ chain: B]

```
>3QWQB        Partners:A
DVPRDLEVVAATPTSLLISWDSGRGSYQYYRITYGETGGNSPVQEFTVPGPVHTATISGLKPGVDYTITVYAVTDHKPHTYHESPISINYRTEID
Seq length: 95

:: SVM Results ::        View in 3D (Astex)
30:1.00 31:2.12 48:0.66 51:2.69 77:0.04
>3QWQB
DVPRDLEVVAATPTSLLISWDSGRGSYQYYRITYGETGGNSPVQEFTVPGPVHTATISGLKPGVDYTITVYAVTDHKPHTYHESPISINYRTEID

:: Ensemble Results ::        View in 3D (Astex)
30:1.00 31:1.00 48:1.00 49:1.00 51:1.00 54:1.00 77:1.00
>3QWQB
DVPRDLEVVAATPTSLLISWDSGRGSYQYYRITYGETGGNSPVQEFTVPGPVHTATISGLKPGVDYTITVYAVTDHKPHTYHESPISINYRTEID
```

Residue ID: 77D, Prediction score: 0.04

Prediction of Hot Spots by Using Structural Neighborhood Properties

Query Results of Job: HS1226098 Cutoff : 0

Job ID: HS1226098
E-mail: test@gmail.com
Query Protein: 3QWQ - B

Job Title: test_3QWQ
Submit Time: 2013-12-26 06:43:12
Partners: 3QWQ - A

Structural Model

Structural Model
☑ Query 3QWQ_B □
☑ Partner 3QWQ_A 🟩

Surface
☑ Query 3QWQ_B □
☐ Partner 3QWQ_A 🟩

Interface
☐ Interface 3QWQ_B 🟦

51:P(2.68) 31:Y(2.11) 30:Q(0.99) 48:F(0.65) 77:D(0.04)

Figure 2. A snapshot of PredHS prediction output. (**A**) SVM results and Ensemble results of a job are listed. Predicted hot spots are colored in red. By default, PredHS predicts interfaces to be hot spots with a default cutoff 0 which is adjustable by the user. Users can put the mouse over a residue to view its residue ID and predicted score. (**B**) Interactive 3D view for a prediction. Predicted hot spots are colored according to their predicted scores. Residues with score higher than zero are shown from light red to red as the score increases.

Recently, we developed an effective structure-based hot spot prediction method, PredHS (22), which integrates novel structural and energetic features with Euclidian and Voronoi neighborhoods in addition to conventionally used properties. Moreover, PredHS uses a two-step hybrid approach to select an optimal subset of features. Based on the selected features, a support vector machine (SVM) classifier and an ensemble model are built for prediction. We have benchmarked PredHS using a set of experimentally verified hot spot residues and an independent dataset. Results show that PredHS significantly outperforms the state-of-the-art methods and indicate that structural neighborhood properties are important determinants of hot spots (22).

Here, we present the PredHS web server, which is an automatized online implementation of the PredHS method. The server allows users to request new predictions for input PDB IDs or structures files provided in PDB format. The resulting predictions can be visualized in an interactive 3D viewer and downloaded as text files.

METHODS

The computational approach used by the PredHS web server consists of three main component processes (Figure 1): (i) feature extraction: to extract a wide variety of sequence, structural and energy features, together with two types of structural neighborhoods; (ii) feature selection:

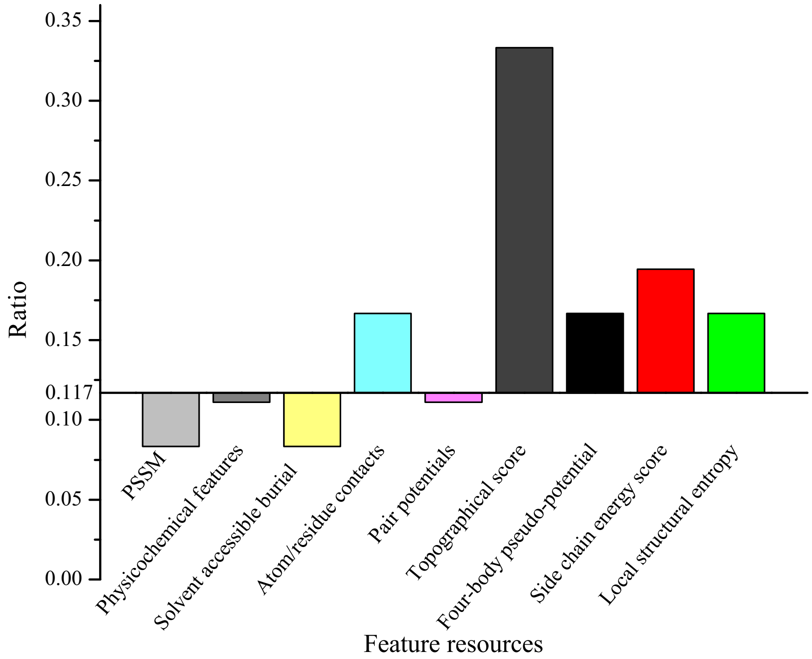


Figure 3. Relative contribution of different feature resources. The y-axes show the ratio of the individual feature resource occurred in the selected optimal features. The reference ratio of 0.117 is the ratio of 38 selected features out of the total number of features.

a two-step feature selection process that combines random forest and a sequential backward elimination and (iii) predictor construction: two predictors (PredHS-SVM and PredHS-Ensemble) are built for identifying hot spots based on the optimally selected features.

Structural neighborhood properties

A total of 108 features are extracted to describe potential hot spot residues. In addition to conventionally used properties, many novel structural and energetic features are also used, including local structural entropy (23), side chain energy score (24), four-body pseudo-potential (25) and topographical score (16,26). Based on these features, we propose a new way to calculate two types of structural neighborhood properties using Euclidean distance and Voronoi diagram.The Euclidean property of a target residue is defined by summing up the values of the properties in the neighborhood. The Euclidean neighborhood is a group of residues located within a sphere of 5 Å defined by the minimum Euclidean distances between each heavy atom of the surrounding residues and each heavy atom of the target residue. The Voronoi diagram (27) is another way to calculate structural neighborhood properties, which partitions a 3D space (a protein structure) into several Voronoi regions, each of which contains a point (heavy atom of a residue). A pair of residues are said to be each other's neighbor when there is at least one pair of their heavy atoms has a Voronoi facet in common. The Voronoi partition is computed by Qhull (28). This definition is based on geometric partitioning rather than the use of an absolute distance cutoff, and hence is considered to be more robust (29).

Two-step feature selection

To remove potentially redundant ones from the whole set 108 features we implement a two-step strategy. The first step is to evaluate the importance of each candidate feature by the mean decrease Gini index (MDGI) with the random forest (RF) package in R (30). A higher MDGI score means the feature is more informative for classifying an interface residue into hot spots and non-hot spots. In PredHS, 77 features with MDGI Z-score larger than 2.5 are selected. In the second step, redundant features are removed by sequential backward elimination (SBE) with 10-fold cross-validation. The SBE algorithm sequentially removes features from the whole feature set till an optimal feature subset is obtained. A feature is removed if its removal maximizes the performance of the predictor. Finally, an optimal set of 38 features is obtained for building prediction models (22).

Prediction models

Two classifiers were built for hot spot prediction: one is PredHS-SVM, which is implemented with LIBSVM package (31) using radial basis function (RBF) as the kernel; the other is an ensemble classifier, PredHS-Ensemble, which is built to handle the problem of imbalance in classification. PredHS-Ensemble uses an ensemble of *n* sub-models that employ an asymmetric bootstrap resampling approach to generate subsets. Each subset contains all of the hot spots and a subset of non-hot spots that is generated using random bootstrap sampling and has the same size as hot spots. The final results are calculated by majority votes among the outputs of the *n* sub-models.

WEB SERVER INTERFACE

Users can upload a file of a protein structure in the PDB format or simply input a PDB code to start a job. The input structure should contain at least two chain identifiers forming an interface. Multiple structures can be submitted in one run. Users could choose to leave their email address or a job title to conveniently retrieve the results. The PredHS server first checks the validity of the input structure, and once confirmed, it progresses to the second step for users to select the query protein and its partners. When the selection is done, users can submit the prediction job by clicking the 'Submit' button.

A typical query takes no more than 30 min to run. For each submitted structure, the server returns two lists of residues and their associated scores to be hot spot, corresponding to PredHS-SVM and PredHS-Ensemble, respectively (Figure 2A). The red residues in the query sequence are predicted hot spots. Users can view the residue ID and its associated score by putting the mouse over the residue. The higher the score is, the more likely a given residue is a hot spot. The results can be downloaded in text or visualized in an interactive 3D viewer AstexViewer (32) by following the 'View in 3D' link. As shown in Figure 2B, predicted hot spots are colored according to their associated scores.

RESULTS

The PredHS web server trains prediction models based on a dataset of 265 experimentally mutated interface residues obtained from ASEdb (3) and the published data of Kortemme and Baker (10), among which 65 are hot spots. To make a fair comparison with other methods, we use an independent test dataset extracted from the BID database that contains alanine-mutation experiments of a different set of 127 interface residues, of which 39 are identified as hot spots.

We calculated a variety of measures to evaluate the predictions:

$$\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN});$$

$$\mathrm{Specificity} = \mathrm{TN}/(\mathrm{FP} + \mathrm{TN});$$

$$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP});$$

$$\mathrm{Accuracy} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN});$$

$$\mathrm{CC} = \frac{(\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}};$$

$$F1 = 2\times\mathrm{Precision}\times\mathrm{Recall}/(\mathrm{Precision} + \mathrm{Recall}).$$

Here TP, FP, TN and FN are true positive, false positive, true negative and false negative counts. We also calculated the area under the receiver operating characteristic (ROC) curve (AUC).

PredHS predicts hot spots using an optimal feature set of 38 features, which are selected from the combination of 108 site features, 108 Euclidean features and 108 Voronoi features with the proposed two-step feature selection approach. These features are calculated based on heterogeneous information resources, including position-specific scoring matrix (PSSM) (33), physicochemical features (34), solvent accessible burial (35), atom/residue contacts (13), pair potentials (36), topographical score (16,26), four-body pseudo-potential (25), side chain energy score (24) and local structural entropy (23). To analyze the relative contribution of different feature resources, we calculated the ratio of each feature resource occurred in the selected optimal set (37). The ratio of 0.117 is used as a reference ratio since it is the ratio of selected 38 optimal features out of the total number of features. From Figure 3, we can see that the topographical score contributes most to the hot spot identification, followed by the side chain energy score, atom/residue contacts, four-body pseudo-potential and local structural entropy.

We used 10-fold cross validation on the training dataset to evaluate the predictive power of structural neighborhood properties and selected optimal features. Five SVM classifiers were built and tested using five groups of features, including site, sequence, Euclidean, Voronoi and optimal features. The sequence features are generated with a sliding window of 21, which includes 10 residues upstream and 10 residues downstream of the target residue in the protein sequence. As shown in Table 1, classifiers with structural neighborhood properties (Euclidean and Voronoi) achieve better performance than those using site and sequence features in terms of AUC score. The classifier with linear sequence neighborhood properties is significantly worse than the others, and thus the sequence features are not included in the combination. PredHS-SVM with the optimal features achieves the best performance, suggesting that the proposed two-step feature selection algorithm can effectively improve the prediction.

Furthermore, we compared PredHS with other five state-of-the-art methods, including Robetta (10), FOLDEF (9), HotPoint (38), KFC2a and KFC2b (15). Each method has a companion web server or a stand-alone software. Results of the independent BID dataset are shown in Table 2. PredHS significantly outperforms the existing methods in the five performance measures (accuracy, specificity, precision, CC and *F*1 score). Although KFC2a has a similar recall value (0.74) to that of PredHS-Ensemble, the specificity (0.74) and precision (0.56) of KFC2a are much lower than that (0.80 and 0.63) of PredHS-Ensemble.

CONCLUSION

The PredHS web server provides an automated platform to predict hot spots from interfaces. In contrast to the approaches based on the recognition of differences in physicochemical properties, evolutionary conservation and solvent accessible area, an advantage of PredHS is that it integrates Euclidian and Voronoi neighborhoods together with a variety of heterogeneous information, including sequence-based, structure-based and energetic features. What's more, PredHS uses a two-step feature selection approach, providing an effective way for selecting an optimal subset of features within a reasonable computational cost, which improves the prediction performance and reduces the risk of over-fitting.

A limitation of PredHS and many other hot spot prediction methods is that they can only identify hot spots from known protein interfaces, which means that the input to

Table 1. Prediction performance comparison of classifiers with different types of features (site, sequence, Euclidean, Voronoi and optimal subset)

Feature types	AUC	Accuracy	Recall	Specificity	Precision	CC	*F*1
Site	0.81	0.82	0.62	0.89	0.66	0.52	0.63
Sequence	0.76	0.80	0.39	0.93	0.77	0.42	0.46
Euclidean	0.82	0.82	0.57	0.91	0.67	0.50	0.60
Voronoi	0.83	0.84	0.60	0.92	0.75	0.57	0.65
Optimal subset (PredHS-SVM)	0.87	0.88	0.75	0.93	0.79	0.69	0.76

Table 2. Prediction performance comparison on the independent BID dataset. Maximum value(s) of each performance measure is(are) highlighted in bold.

Methods	Accuracy	Recall	Specificity	Precision	CC	*F*1
Robetta	0.70	0.33	0.86	0.52	0.23	0.41
FOLDEF	0.68	0.26	0.87	0.48	0.16	0.33
HotPoint	0.69	0.59	0.74	0.5	0.31	0.54
KFC2a	0.74	**0.74**	0.74	0.56	0.41	0.64
KFC2b	0.79	0.59	0.87	0.68	0.47	0.63
PredHS-SVM	**0.83**	0.59	**0.93**	**0.79**	**0.57**	**0.68**
PredHS-Ensemble	0.79	**0.74**	0.80	0.63	0.53	**0.68**

these methods should be protein complexes forming interfaces other than monomers. We plan to improve PredHS by using structural alignment methods to detect hot spots from predicted interfaces, and thus make monomer input possible.

PredHS has been in service for >10 months and it is under continuous improvement. We hope PredHS can be applied to a wide range of hot spot identification and further functional analysis and so to provide a practical tool for biologists.

ACKNOWLEDGMENT

L.D. thanks for support from postdoctoral research station of Central South University.

FUNDING

National Natural Science Foundation of China [61309010, 61173118, 61272380]; China 863 Program [2012AA020403]; Specialized Research Fund for the Doctoral Program of Higher Education of China [20130162120073]. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

1. Clackson,T. and Wells,J.A. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science (New York, N.Y.)*, **267**, 383–386.
2. Wells,J.A. and McClendon,C.L. (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, **450**, 1001–1009.
3. Thorn,K.S. and Bogan,A.A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284–285.
4. Fischer,T.B., Arunachalam,K.V., Bailey,D., Mangual,V., Bakhru,S., Russo,R., Huang,D., Paczkowski,M., Lalchandani,V., Ramachandra,C. *et al.* (2003) The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics*, **19**, 1453–1454.
5. Massova,I. and Kollman,P.A. (1999) Computational alanine scanning to probe protein–protein interactions: a novel approach to evaluate binding free energies. *J. Am. Chem. Soc.*, **121**, 8133–8143.
6. Grosdidier,S. and Fernandez-Recio,J. (2008) Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics*, **9**, 447.
7. Brenke,R., Kozakov,D., Chuang,G.Y., Beglov,D., Hall,D., Landon,M.R., Mattos,C. and Vajda,S. (2009) Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics*, **25**, 621–627.
8. Huo,S., Massova,I. and Kollman,P.A. (2002) Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *J. Comput. Chem.*, **23**, 15–27.
9. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
10. Kortemme,T. and Baker,D. (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 14116–14121.
11. Ofran,Y. and Rost,B. (2007) Protein-protein interaction hot spots carved into sequences. *PLoS Comput. Biol.*, **3**, e119.
12. Darnell,S.J., Page,D. and Mitchell,J.C. (2007) An automated decision-tree approach to predicting protein interaction hot spots. *Proteins*, **68**, 813–823.
13. Cho,K.I., Kim,D. and Lee,D. (2009) A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res.*, **37**, 2672–2687.
14. Xia,J.F., Zhao,X.M., Song,J. and Huang,D.S. (2010) APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics*, **11**, 174.
15. Zhu,X. and Mitchell,J.C. (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins*, **79**, 2671–2683.
16. Assi,S.A., Tanaka,T., Rabbitts,T.H. and Fernandez-Fuentes,N. (2010) PCRPi: presaging critical residues in protein interfaces, a new computational tool to chart hot spots in protein interfaces. *Nucleic Acids Res.*, **38**, e86.
17. Segura,J. and Fernandez-Fuentes,N. (2011) PCRPi-DB: a database of computationally annotated hot spots in protein interfaces. *Nucleic Acids Res.*, **39**, D755–760.
18. Tuncbag,N., Gursoy,A. and Keskin,O. (2009) Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, **25**, 1513–1520.
19. Wang,L., Liu,Z.P., Zhang,X.S. and Chen,L. (2012) Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Eng. Des. Sel.*, **25**, 119–126.

20. Cukuroglu,E., Gursoy,A. and Keskin,O. (2012) HotRegion: a database of predicted hot spot clusters. *Nucleic Acids Res.*, **40**, D829–D833.
21. Guney,E., Tuncbag,N., Keskin,O. and Gursoy,A. (2008) HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res.*, **36**, D662–D666.
22. Deng,L., Guan,J., Wei,X., Yi,Y., Zhang,Q.C. and Zhou,S. (2013) Boosting prediction performance of protein-protein interaction hot spots by using structural neighborhood properties. *J. Comput. Biol.*, **20**, 878–891.
23. Chan,C.H., Liang,H.K., Hsiao,N.W., Ko,M.T., Lyu,P.C. and Hwang,J.K. (2004) Relationship between local structural entropy and protein thermostability. *Proteins*, **57**, 684–691.
24. Liang,S. and Grishin,N.V. (2004) Effective scoring function for protein sequence design. *Proteins*, **54**, 271–281.
25. Krishnamoorthy,B. and Tropsha,A. (2003) Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, **19**, 1540–1548.
26. Levy,E.D. (2007) PiQSi: protein quaternary structure investigation. *Structure (London, England:1993)*, **15**, 1364–1367.
27. Okabe,A., Boots,B. and Sugihara,K. (1992) *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley & Sons, Inc, Hoboken, New Jersey, United States, 65-115.
28. Barber,C.B., Dobkin,D.P. and Huhdanpaa,H. (1996) The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, **22**, 469–483.
29. Segura,J., Jones,P.F. and Fernandez-Fuentes,N. (2011) Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. *BMC Bioinformatics*, **12**, 352.
30. Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.
31. Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 1–27.
32. Hartshorn,M.J. (2002) AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aided Mol. Des.*, **16**, 871–881.
33. Deng,L., Guan,J., Dong,Q. and Zhou,S. (2009) Prediction of protein-protein interaction sites using an ensemble method. *BMC Bioinformatics*, **10**, 426.
34. Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.
35. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
36. Keskin,O., Bahar,I., Badretdinov,A.Y., Ptitsyn,O.B. and Jernigan,R.L. (1998) Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci.*, **7**, 2578–2586.
37. Hu,L., Cui,W., He,Z., Shi,X., Feng,K., Ma,B. and Cai,Y.D. (2012) Cooperativity among short amyloid stretches in long amyloidogenic sequences. *PloS One*, **7**, e39369.
38. Tuncbag,N., Keskin,O. and Gursoy,A. (2010) HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res.*, **38**, W402–W406.